Hybrid Simulator for Capturing Dynamics of Synthetic Populations

Marija Kukic, Salim Benchelabi and Michel Bierlaire

***Abstract*— This paper presents a novel hybrid framework for generating and updating a synthetic population. We call it hybrid because it combines model-based and data-driven approaches. Existing generators produce a snapshot of synthetic data that becomes outdated over time, requiring complete re-generation using the newest datasets for updates. By leveraging regularly collected data, we propose a method that provides up-to-date synthetic populations at any given moment without using complete re-generation. Our approach generates a baseline synthetic population once, using the Markov Chain Monte Carlo simulation, and projects it over time. In scenarios where disaggregated real data are unavailable, we project the synthetic sample by simulating life-changing events. When new disaggregated real data become available, we calibrate the projected sample using resampling to account for data collection biases and projection errors. We implement and test our approach on 2010, 2015, and 2021 Swiss mobility and transport micro-census data. To generate the baseline sample we use data from 2010 and project it to 2021. We compare the projections of our hybrid approach to existing methods, namely dynamic projection and resampling. The results demonstrate that the synthetic sample generated by the hybrid approach improves the fit to the real data compared to the dynamic projection, and improves heterogeneity compared to the resampling.**

I. INTRODUCTION

In the transportation field, activity-based models (ABM) are used to analyze the travel behavior of individuals, in order to forecast the demand and impacts of various policies [1], [2]. To ensure the robustness of these models across different scenarios, diverse and unbiased datasets are needed for testing and calibration. However, obtaining such real datasets is challenging due to cost and privacy constraints. Synthetic data offer a solution by combining various data sources to generate data that meet specific requirements and reduce bias. Also, ABMs typically require synthetic data to simulate hypothetical scenarios, such as examining the impact of measures or policies on people's behavior or predicting responses of a specific group to future changes.

Existing synthetic generators use aggregated or disaggregated real data as input, from one or several sources, to create synthetic samples that replicate the distributions of real data at a specific time, i.e., synthetic snapshots. This means that once the initial synthetic population is generated, any changes in the reference data cannot be integrated into the synthetic population. However, the real population evolves through different demographic events, e.g., changing the marital or employment status, giving birth, or passing away. Consequently, the synthetic snapshot might become obsolete over time as it no longer reflects the reality [3].

M. Kukic, S. Benchelabi, and M. Bierlaire are with the Ecole Polytechnique Fédérale de Lausanne (EPFL), Transport and Mobility Laboratory, Lausanne 1015, Switzerland. {marija.kukic, salim.benchelabi, michel.bierlaire}@epfl.ch.

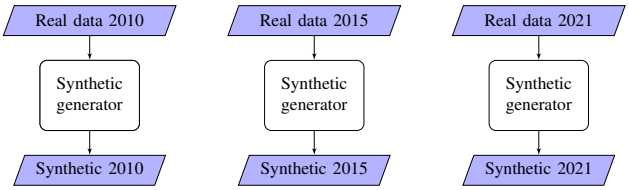

Fig. 1: The independent generation

As illustrated in Fig. 1, the current literature suggests that obtaining up-to-date synthetic data requires re-generating a new synthetic sample based on the latest real data. This method, called independent generation, does not track changes between past, present, and future data and imposes no relationships between successive iterations of the same synthetic population across time. This makes the generation unnecessarily complicated and costly. Firstly, adapting a synthetic generator to a new dataset requires repeating the same, usually highly problem-specific, procedure from before. Secondly, most synthetic generators require disaggregated real data as input to generate a high-quality synthetic sample. Although regularly collected data, such as census, open many possibilities for enhancing the quality of synthetic data, to the best of our knowledge, there is no methodology that exploits all past data from previous surveys.

Some authors have proposed to project the synthetic snapshot to capture changes over time, i.e., dynamics, and test future hypothetical scenarios [4], [5], [6]. All of these methods contain two steps: generation, i.e., initialization of the baseline synthetic population, and its projection over time, i.e., evolution. Since the quality of the projection is highly dependent on the quality of the baseline sample, the synthetic sample should be thoroughly validated before performing the projection.

Once the synthetic sample is generated, the projection step is usually performed using dynamic projection [4], [5] or resampling [6]. Dynamic projection simulates specific events and their impact on desired synthetic attributes, e.g., socio-demographic characteristics of individuals. On the other hand, resampling adjusts the synthetic data to match the marginals of the newly obtained real dataset by randomly adding or deleting observations. However, both of these approaches suffer from certain limitations.

Usually, the census data collection initiative does not track the same individuals over years, resulting in different sampling biases in each sample. Because of this, the dynamic projection method inherently propagates bias from the baseline population, year by year, which decreases the

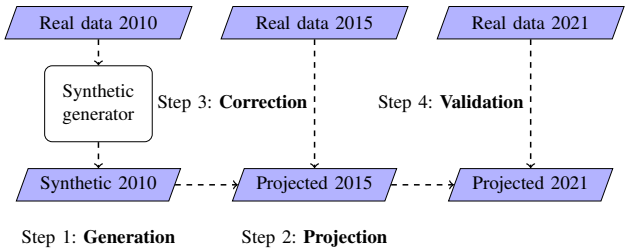

Fig. 2: Hybrid Simulator for Capturing Dynamics of Synthetic Populations

goodness of fit between the projected sample and the most recent real data. In addition, the projection rates are usually derived under the assumption that the population follows similar trends over time, which is not always true. For instance, unexpected major events, e.g., Covid-19, can lead to non-representative projected samples compared to the real data. On the other hand, resampling randomly selects the individuals to duplicate or delete, which might result in always choosing people with the same characteristics, hence reducing the synthetic population heterogeneity.

Once the sample is projected to a specific point in time, authors validate their results by comparing synthetic distributions against the real ones. However, the limitations of the projection methods become more significant the further away in time we try to project, resulting in an accumulated propagated error that could impact the representativity of the projected sample. Thus, the validation should not be performed only once, but regularly.

In this paper, we propose a hybrid simulator that consists of four parts: generation, dynamic projection, resampling (i.e., correction), and validation, as shown in Fig. 2. Our approach is hybrid in the sense that it is both model-based, due to dynamic projection, and data-driven, due to resampling. We use dynamic projection to simulate life events when disaggregated real data are not yet available, and once the new real data are released, we correct projection errors by adapting the projected marginals according to the new data. In contrast to other projection methods that only use the past to generate the future, we also exploit the information from the present to improve the past.

In our framework, the sample is generated once, and then regularly updated and validated, which has several benefits. Firstly, our method maintains the synthetic data up-to-date and enriches the resulting synthetic sample by taking into account all available census data. Secondly, due to correction and validation steps, the projection method is robust enough to deal with unusual events, and the resulting sample is less impacted by errors in the generation step, over long projection horizons. Finally, updating already generated samples can potentially reduce the data collection cost, since we can use fewer data for future updates.

The remainder of this paper is organized as follows: Section II covers a review of the existing methods that combine different generation and projection methods. In Section III, we formally describe each part of the hybrid simulator. Finally, in Sections IV and V we present the results of comparisons between different methods, summarize the research contributions, and present some ideas for future research.

II. RELATED WORK

A lot of research effort has been invested in improving the algorithms for generating synthetic populations. These algorithms try to replicate the distributions of the individual, i.e., one-level generation, and household attributes, i.e., multi-level generation, while preserving the representativity and realism of the real data. An extensive literature review on the existing individuals and household generation algorithms has been provided by [7] and [8], respectively. Based on these studies, both one-level and multi-level generation can be categorized into three groups: synthetic reconstruction, e.g., iterative proportional fitting (IPF), iterative proportional updating (IPU); combinatorial optimization, and statistical learning, e.g., Markov chain Monte Carlo simulation (MCMC), machine learning methods (ML). All these methods can be appropriate in specific situations and the performance, e.g., realism, representativity, computational time, can differ depending on various factors, such as sample size, number of attributes, structure of the input dataset, etc. As suggested in [8], there is no consensus on which synthetic generation method is the best in the general case and the choice of the algorithm depends on the users' needs and available resources.

Consequently, algorithms based on dynamic projection [4], [5], [9], static projection [10], and resampling [6] have emerged. These works consider a similar set of attributes: age and gender at the level of individuals, and household size and type at the level of the households. Dynamic projection simulates the effects of death, birth, couple formation, couple dissolution, and leaving home on the age and gender of individuals. Static projection, on the other hand, consists of applying a reconstruction method on the synthetic sample to modify it according to the latest real aggregated data. Lastly, resampling randomly adds and removes individuals based on discrepancies between the marginals of the real data and the marginals of the projected population. None of the cited papers propose a combination of projection methods.

As shown in Table I, these papers can be categorized based on methods used for generation and projection steps. Interestingly, most of them use synthetic reconstruction for generating the baseline population. Although statistical learning methods have shown better results [11], none of the aforementioned papers use them in the generation step. This might be because most of the publicly available population synthesizers use IPF as the core algorithm [12]. The generator can be chosen arbitrarily when the projection is limited to a few attributes. However, it has been shown that IPF struggles to deliver acceptable results when working with high-dimensional datasets [13]. Since some algorithms cannot maintain generation quality with increasing scale, when performing projection with more attributes, the choice of the synthetic generator is one of the crucial decisions.

TABLE I: The review of generation and projection algorithms

	Dynamic projection	**Static projection**	**Resampling**
Synthetic reconstruction	*Baseline Synthesis and Microsimulation of Life-stage Transitions within an Agent-based Integrated Urban Model* Fatmi et al. **2017**	*An Open-Source Model for Projecting Small Area Demographic and Land-Use Change* Lomax et al. **2022**	*A synthetic population for agent-based modelling in Canada* Prédhumeau et al. **2023**
Combinatorial optimisation	*Generating a Dynamic Synthetic Population Using an Age-Structured Two-Sex Model for Households Dynamics* Namazi-Rad et el. **2014**	**X**	**X**
Statistical learning	**Hybrid Simulator for Capturing Dynamics of Synthetic Populations Model-based**	**X**	**Hybrid Simulator for Capturing Dynamics of Synthetic Populations Data-driven**

III. METHODOLOGY

In this section, we provide details on each part of the developed pipeline, where the output of each phase is the input to the next one. Although this methodology can be applied using any dataset, in this paper we focus on the Swiss Mobility and Transport microcensus data (MTMC) from 2010, 2015, and 2021, provided by the Swiss Federal Statistical Office (BFS) [14].

A. Generating and validating the baseline synthetic sample

The baseline sample is generated only once using the MCMC simulation, more precisely Gibbs sampling (GS) [11]. Inspired by the methodology introduced in [15], we propose a simplified adaptation of their algorithm for household generation and use it to generate individuals instead. The synthetic sample contains vectors X composed of discrete random variables that represent individual characteristics. In the current version of the framework, we generate age X_{a}, employment X_{e}, and gender X_{g}. The possible values of the chosen attributes are described in Section IV-A. The goal of GS is to reproduce the multivariate joint distribution of these attributes denoted by $\pi(X)$, by randomly drawing at each iteration a value of one attribute conditioned on all others. If N is the number of attributes that we generate, and $\mathcal{A} = \{X_{\mathrm{i}} | \mathrm{i} \in \{1, ..., N\}\}$ is the set of random variables that describes the attributes of a certain individual, then the generated value x_{k} of a randomly selected variable $X_{\mathrm{k}} \in \mathcal{A}$ is drawn from the conditional distribution $\pi(X_{\mathrm{k}} | \mathcal{A} \setminus \{X_{\mathrm{k}}\})$ for an a priori fixed realization of the random variables defined by $\mathcal{A} \setminus \{X_{\mathrm{k}}\}$, i.e., for $X_{\mathrm{i}} = x_{\mathrm{i}}, \forall \mathrm{i} \in \{1, ..., N\} \setminus \{\mathrm{k}\}$. These conditional distributions are derived from data and provided as input to the algorithm. Simulated chains of each attribute should converge to a unique joint distribution within the specified number of draws. To monitor the convergence and to identify if the algorithm has reached the stationary state, we compute the potential scale reduction factor and the effective sample size. For further explanation on these metrics, interested readers are referred to [16].

Once the baseline sample is generated, the joint distribution of simulated attributes is compared against the real data used as a reference. This step is essential to analyze what percentage of the error is propagated in the projection step. To validate each characteristic separately, i.e., marginal distributions, we use standardized root mean squared error (SRMSE) as shown in (1) [13], [17].

$$\mathrm{SRMSE} = \frac{\sqrt{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} \frac{(\pi^{\mathrm{synth}}_{i,\ldots,j} - \pi^{\mathrm{real}}_{i,\ldots,j})^2}{N_{\mathrm{cnt}}}}}{\sum_{i=1}^{m} \cdots \sum_{j=1}^{n} \frac{\pi_{i,\ldots,j}}{N_{\mathrm{cnt}}}} \tag{1}$$

Here, π^{synth} and π^{real} represent the frequency count of each unique combination of attributes $(i, ..., j)$, in the real and synthetic samples, respectively, where $(m, ..., n)$ are the numbers of possible categories of these attributes. N_{cnt} denotes the total number of unique combinations of values for attributes $(i, ..., j)$. In other words, we calculate the occurrence of unique values for each combination of arbitrarily chosen real and corresponding synthetic columns and compare them. We additionally compute the SRMSE to systematically test all possible combinations of all columns on different aggregation levels, using the statistical framework proposed by [18]. With an aggregation level, we specify the number of columns that are jointly assessed. Namely, if N_{v} is the set of n_{v} columns in the dataset, for a specified aggregation level $a \in \{1, 2, 3\}$, we calculate SRMSE for all possible $\binom{n_{\mathrm{v}}}{a}$ frequency lists. The final result is the average of all previously calculated SRMSE scores. Note that although we present only the SRMSE in our results, we also validate our results using other statistics available in the framework such as mean absolute error (MAE), coefficient of determination (R^2), and root mean square error (RMSE).

B. Dynamic projection

Let t_0 be the year when the baseline synthetic sample is generated, and t_{end}, where $t_{\mathrm{end}} > t_0$, the year to which we want to project. For each year t_{n}, where $t_0 < t_{\mathrm{n}} < t_{\mathrm{end}}$, we simulate the effects of births, deaths, and migrations on the attributes of the synthetic population. As shown in Algorithm 1, we update the synthetic sample from time step $t_{\mathrm{n\text{-}1}}$ to time step t_{n} by projecting the age, gender, and employment of synthetic individuals. The output of this method is an updated synthetic sample that we call a projected sample. To simulate each event, we use rates of people of a certain age and gender that gave birth, died, immigrated, or emigrated. The rates we use are provided for each year t_n at the aggregated level by the Swiss Federal Statistical Office [14]. To simulate births, we use data about the number of births based on the mother's age and the number of women within each age group. We compute fertility rates for each age class by dividing the number of births by the number of women. For each female individual, we randomly simulate giving birth using the probability with respect to the women's age. To compute mortality

rates, we use the number of deaths per age and gender. For each individual in the dataset, we randomly remove an individual from the sample using the corresponding mortality rate as the binomial probability. Furthermore, we simulate immigration and emigration by adding or removing people, using statistical records of net migration per year, age and gender. The net migration presents the difference between the number of immigrants and the number of emigrants. The rates for simulation are computed similarly. Finally, we deterministically assign the employment status 'retired' if the person is above 65, and 'under 15' if the person is under 15, since we do not have access to the employment rates. For other age categories, we draw the employment status conditioned on the given age and gender of the person, using the baseline sample as a reference.

Algorithm 1 Dynamic projection

```
function DYNAMIC_PROJECTION(synthetic_sample, t_0, t_end)
    predictive_sample = synthetic_sample
    for i = t_0 to t_end do
        increment_age(predictive_sample);
        add_children(predictive_sample, i);            ▷ Birth rates
        remove_individuals(predictive_sample, i);      ▷ Death rates
        add_individuals(predictive_sample, i);         ▷ Migration rates
        remove_individuals(predictive_sample,i);
    end for
    draw_employment(predictive_sample);
end function                                           ▷ Return the updated sample
```

C. Resampling

Let t_0 and t_c, where $t_c > t_0$, be the years when two consecutive census surveys are performed. We assume that the two obtained datasets share a portion of respondents with similar characteristics. Therefore, there is a logical shift in the distributions that should be captured by dynamic projection. Since the census data do not track the same individuals over the years, we cannot identify what portion of the data is the same, leading to different biases in each census dataset. Thus, by projecting we propagate the sampling bias from the year t_0. To correct the propagated bias and projection errors, at year t_c, when the new census disaggregated data become available, we perform resampling as shown in Algorithm 2.

We update the marginals of the projected sample from t_0 to t_c, based on the age marginals of the new real census data released at year t_c. Comparing the frequency counts of the age categories between these two samples, we randomly duplicate or remove individuals from the projected sample for each age group, in order to achieve a better fit to the most recent real data. We define a specific threshold whose smaller value indicates a better fit. Given that most of the simulated events, such as giving birth, death, migration, etc., are age dependent, we decide to resample only based on this attribute. Since age is correlated with most of the attributes, changing it can implicitly impact other highly correlated variables. For instance, adding more students will consequently increase the number of individuals who are in education. Since we recycle people from the synthetic dataset, we have to make sure that the sub-distributions are also well replicated in the baseline synthetic dataset. Otherwise, if there are some illogically generated individuals, e.g., people under 15 who are retired, we risk duplicating them by resampling which impacts the quality of the resulting sample. The resampled dataset created at t_c is used as a new baseline to continue the dynamic projection until the final year t_{end}.

Algorithm 2 Resampling procedure

```
function RESAMPLE(a,b,num,threshold)
    a - array of frequency counts per each age category in reference sample
    b - array of frequency counts per each age category in projected sample
    num - total number of age categories
    for i = 1 to num do
        if abs(a[i] − b[i]) > threshold then
            if (a[i] − b[i]) < 0 then
                nb_of_observation = abs(a[i] − b[i])
                for j = 1 to nb_of_observation do
                    randomly sample a person of the age i
                    remove a selected person from the projected sample
                end for
            else
                nb_of_observation = abs(a[i] − b[i])
                for j = 1 to nb_of_observation do
                    randomly sample a person of the age i
                    add the selected person to the projected sample
                end for
            end if
        end if
    end for
end function                                           ▷ Return the updated sample
```

IV. RESULTS

In this section, we implement and verify our hybrid methodology in a case study of Switzerland's population, using MTMC data from 2010, 2015, and 2021. Initially, the synthetic sample from 2010 is generated and validated against the real data from 2010. Subsequently, the synthetic sample from 2010 is projected to 2015 using dynamic projection. Then, we compare the age marginals of the projected sample from 2010 to 2015 with the age marginals of the real MTMC 2015 data. Based on this comparison, we apply resampling on the projected sample in an attempt to improve the fit to the real marginals. Finally, the corrected sample is projected from 2015 to 2021 and validated against the real data from 2021. In order to analyze the stability of the complete method, we run several simulations for both the generation and the projection steps and perform bootstrapping to calculate confidence intervals. The validation is conducted by comparing the marginals and sub-distributions of different generated samples against the real data. In addition, SRMSE scores are reported at three different aggregation levels, i.e., the first, second, and third order. The first order indicates the fit of the marginal distributions, while the second and third-order statistics provide insights into the replication of sub-distributions. We either report the SRMSE value for a particular attribute or a certain combination of them at a specific aggregation level. Finally, the mean of SRMSE is computed for all attributes, with a lower score indicating a better fit.

A. Data description

The MTMC data, usually collected every 5 years, consist of two datasets. One contains information about the sociodemographic attributes of all individuals in the surveyed households, e.g., age and gender, while the other contains information about households, e.g., size, type, number of cars, as well as additional information about the survey respondent, e.g., employment. Since data are collected following similar procedures, all datasets are characterized by a similar set of attributes. In this paper, we focus on the age ([6-99]), employment ('employed', 'unemployed', 'education', 'retired', 'under 15'), and gender ('male', 'female') of an individual in the household.

TABLE II: Data description

	Sample size Original	Sample size Preprocessing	Data loss
MTMC 2010	62,903	62,868	0,06%
MTMC 2015	57,070	57,053	0,03%
MTMC 2021	55,018	54,986	0,06%

Compared to the original dataset, we remove individuals with missing information about any of the attributes. The differences between the original and pre-processed data are shown in Table II. For employment, we aggregate categories: 'self-employed', 'working in a company', 'employee', and 'apprentice' into the group 'employed', and categories 'unemployed', 'unable to work', 'housewife/husband', and 'non-working adult' into the group 'unemployed'. The rest of the categories remain the same. The MTMC dataset provides weights to correct sampling biases, which we have applied in all of our experiments. To validate the generation step, we discretize age further into different categories: '$<$15', '15-17', '18-23', '24-43', '44-64','$\geq$ 65'. Finally, to validate the projection step, we discretize the age with a step of 5 years, to match the time gap between two successive surveys.

B. Evaluation of the baseline synthetic sample

In Fig. 3, we see that the synthetic marginals of each attribute fit the marginals of the real sample used as a reference. Small confidence intervals indicate that the results are stable over several iterations. Some categories of attributes are perfectly correlated, e.g., people under 15 always have employment status 'under 15', and people above 65 are always retired. Therefore, we can simplify the process by deterministically assigning these categories and excluding them from the stochastic generation. As illustrated in Table III, the low scores of SRMSE indicate that the sub-distributions are also well replicated.

C. Comparison of hybrid simulator and state of the art methods - dynamic projection and resampling

In this section, we compare existing dynamic projection (see Section III-B) and resampling (see Section III-C) with our hybrid simulator (see Section IV) that combines both

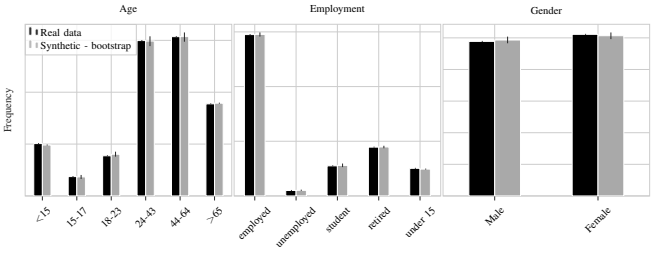

Fig. 3: The comparison of the marginal distributions - synthetic and real sample from 2010

TABLE III: SRMSE - real and generated sample from 2010

	Age	Employment	Gender
First order	$4.45 \cdot 10^{-2}$	$8.76 \cdot 10^{-3}$	$4.69 \cdot 10^{-3}$
	Age, Employment	**Employment, Gender**	**Age, Gender**
Second order	$7.28 \cdot 10^{-2}$	$9.40 \cdot 10^{-3}$	$5.71 \cdot 10^{-3}$
	Age, Employment, Gender		
Third order	$2 \cdot 10^{-4}$		

of them. In order to compare the dynamic projection and hybrid approach we analyze two scenarios, where we project a baseline synthetic sample from 2010 to 2021 using both methods, and compare the obtained distributions against the real data from 2021.

On the left-hand side of Fig. 4, we illustrate the results of dynamic projection, and on the right-hand side, the results of the dynamic projection with resampling, from 2010 to 2015. We notice that using only the dynamic projection yields a worse fit. This comes from the fact that the real samples from 2010 and 2015 do not follow the same age distribution, which results in the propagation of the bias as we project. For example, we observe that the projected sample from 2010 has an over-represented category of children below ten years, compared to the real sample from 2015. This means that this category was over-sampled in the baseline sample.

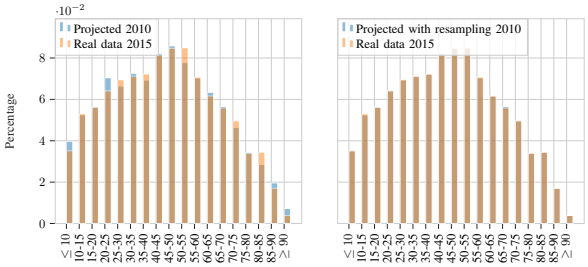

Fig. 4: Comparison of age marginal distributions - dynamic projection (left) and dynamic projection with resampling (right)

In Table IV, we compare the first-order SRMSE of different projected samples over different periods against the real data. The sample projected from 2015 to 2021 is closer to the real data from 2021 than the sample projected from 2010 to 2021. This indicates that using a more recent baseline sample for projection yields a closer fit to the latest data, considering that the error increases as we project over a longer period.

Looking at the projection over a longer time horizon, the hybrid approach gives better results than the dynamic projection from 2010 to 2021. This is expected as the main idea of a hybrid approach is to decrease the projection error by resampling every few years in an attempt to help reduce the bias and improve the accuracy. The age attribute shows the most significant differences since it is used for resampling. However, the correction of age implicitly improves the fit of other correlated attributes.

On the other hand, the projection from 2015 to 2021, i.e., re-generation, gives better results based on the SRMSE than the hybrid approach. This suggests that for smaller projection horizons, there might be no need to perform the resampling step. However, for a larger problem scale, e.g., generating more attributes, the re-generation could suffer from the curse of dimensionality, which can hinder the generator's efficiency and the accuracy of the results. Additionally, it is worth noting that some synthetic generators rely on a complete disaggregation of the real data in order to be able to mimic it, whereas the resampling procedure requires only the marginals. This makes the hybrid approach more resilient with respect to problems related to data availability.

We also test a resampling method from 2010 to 2021, where we adjust marginals annually to achieve a perfect fit of age according to real data, starting from modifying a baseline synthetic sample. While the resampling demonstrates an almost perfect fit for age marginals (refer to Table IV), it leads to a lack of heterogeneity in the generated sample over long projection horizons, resulting in very similar individuals. This phenomenon is evident through the third-order SRMSE score of the resampling (0.33), as compared to the hybrid approach (0.20), which indicates a lower representativity of the sub-distributions. Note that this difference might be more significant when generating a greater number of attributes or attributes that are less correlated with age.

TABLE IV: First order SRMSE - Comparison of different projection scenarios against the real sample 2021

	Age $\cdot 10^{-2}$	Employment $\cdot 10^{-2}$	Gender $\cdot 10^{-2}$	Average All attributes $\cdot 10^{-2}$
Dynamic projection 2015 - 2021	5.76	3.71	0.48	3.31
Hybrid simulator projection 2010 - 2021	7.35	5.26	0.61	4.41
Dynamic projection 2010 - 2021	8.28	7.13	0.67	5.36
Resampling 2010 - 2021	1.69	4.02	1.76	2.49

V. CONCLUSION

In this paper, we present a hybrid framework for generating and maintaining synthetic samples. To the best of our knowledge, this is the first attempt to update synthetic samples by integrating new data without re-generating the entire sample. We compared our approach with existing projection methods using a baseline sample of synthetic individuals. Our results demonstrate that by combining dynamic projection and resampling, we can achieve a better fit between the synthetic data and the most recent real sample compared to the existing techniques when projecting far into the future.

In the future work, we aim to investigate the influence of different factors on the accuracy and the computational efficiency of the re-generation and the hybrid approaches. To establish a general framework, it is of paramount importance to test the scalability of the methods with respect to the number of generated attributes since the complete re-generation method could suffer from the curse of dimensionality. Finally, we aim to test and compare the performance of the two approaches on smaller and sparser datasets.

REFERENCES

[1] J. Castiglione, M. Bradley, and J. Gliebe, *Activity-Based Travel Demand Models: A Primer*, T. R. Board, N. A. of Sciences Engineering, and Medicine, Eds. Washington, DC: The National Academies Press, 2014.

[2] K. W. Axhausen, "Activity-based modelling: Research directions and possibilities," 2000.

[3] F. Gargiulo, S. Ternes, S. Huet, and G. Deffuant, "An Iterative Approach for Generating Statistically Realistic Populations of Households," *PLoS ONE*, vol. 5, no. 1, p. e8828, Jan. 2010.

[4] N. Geard, J. M. McCaw, A. Dorin, K. B. Korb, and J. McVernon, "Synthetic population dynamics: A model of household demography," *Journal of Artificial Societies and Social Simulation*, vol. 16, no. 1, p. 8, 2013.

[5] M.-R. Namazi-Rad, P. Mokhtarian, and P. Perez, "Generating a dynamic synthetic population – using an age-structured two-sex model for household dynamics," *PLOS ONE*, vol. 9, no. 4, pp. 1–16, 04 2014.

[6] M. Prédhumeau and E. Manley, "A synthetic population for agent based modelling in Canada," *Scientific Data*, vol. 10, 03 2023.

[7] D. F. Miranda, "Reviewing synthetic population generation for transportation models over the decades," 2019.

[8] B. F. Yaméogo, P. Gastineau, P. Hankach, and P.-O. Vandanjon, "Comparing methods for generating a two-layered synthetic population," *Transportation Research Record*, vol. 2675, no. 1, pp. 136–147, 2021.

[9] M. Rahman Fatmi and M. Ahsanul Habib, "Baseline synthesis and microsimulation of life-stage transitions within an agent-based integrated urban model," *Procedia Computer Science*, vol. 109, pp. 608–615, 2017.

[10] N. Lomax, A. Smith, L. Archer, A. Ford, and J. Virgo, "An open-source model for projecting small area demographic and land-use change," *Geographical Analysis*, vol. 54, 02 2022.

[11] B. Farooq, M. Bierlaire, R. Hurtubia, and G. Flötteröd, "Simulation based population synthesis," *Transportation Research Part B: Methodological*, vol. 58, 12 2013.

[12] M. Templ, B. Meindl, A. Kowarik, and O. Dupriez, "Simulation of synthetic complex data: The r package simpop," *Journal of Statistical Software*, vol. 79, no. 10, p. 1–38, 2017.

[13] Y. Zhu and J. Ferreira, "Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2429, no. 1, pp. 168–177, Jan. 2014.

[14] Office Fédéral de la Statistique, *Comportement de la population en matière de mobilité*. Neuchâtel: Bundesamt für Statistik (BFS), May, Jan, Apr 2012, 2018, 2023.

[15] M. Kukic and M. Bierlaire, "Divide-and-conquer one-step simulator for the generation of synthetic households," Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Technical Report TRANSP-OR 230408, 2023.

[16] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis (3rd ed.)*. A Chapman and Hall Book,CRC Press, London, 2013.

[17] S. Garrido, S. S. Borysov, F. C. Pereira, and J. Rich, "Prediction of rare feature combinations in population synthesis: Application of deep generative modelling," 2019.

[18] G. Lederrey, T. Hillel, and M. Bierlaire, "Datgan: Integrating expert knowledge into deep learning for synthetic tabular data," 2022.